Predictive Maintenance for Industrial Equipment

Neysha Pagán Vargas,
Charles Hohl, Monica Fam

Agenda

Introduction

- Overview of Predictive Maintenance
- Objectives of the Analysis

Feature Selection & Data Preparation

- Features and Targets
- Data Splitting and Scaling

Model Training & Predictions

- Classification with Random Forest
- Regression with Linear Regression

Model Evaluation

- Regression Metrics: MSE & R^2
- Feature Importance Analysis

Anomaly Detection with Clustering

- Initial Clustering Results
- Silhouette Score and Davies-Bouldin Score

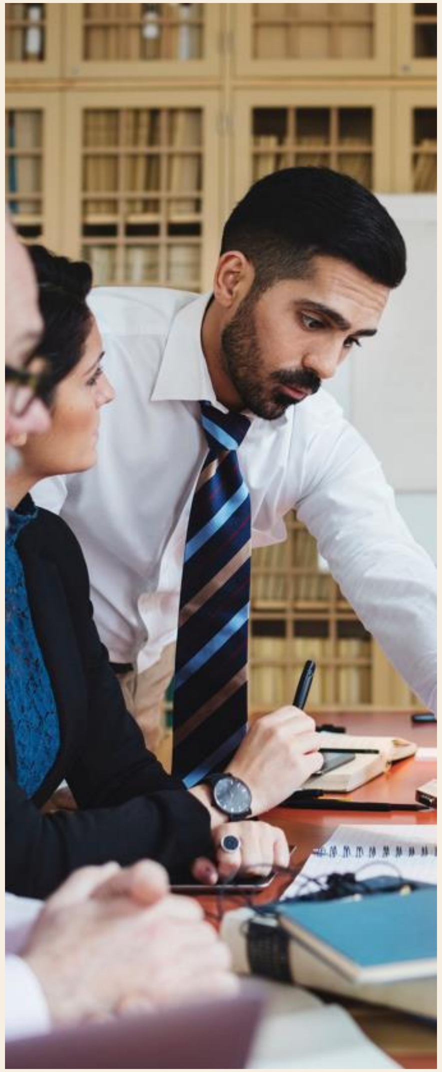
Improving Clustering with PCA

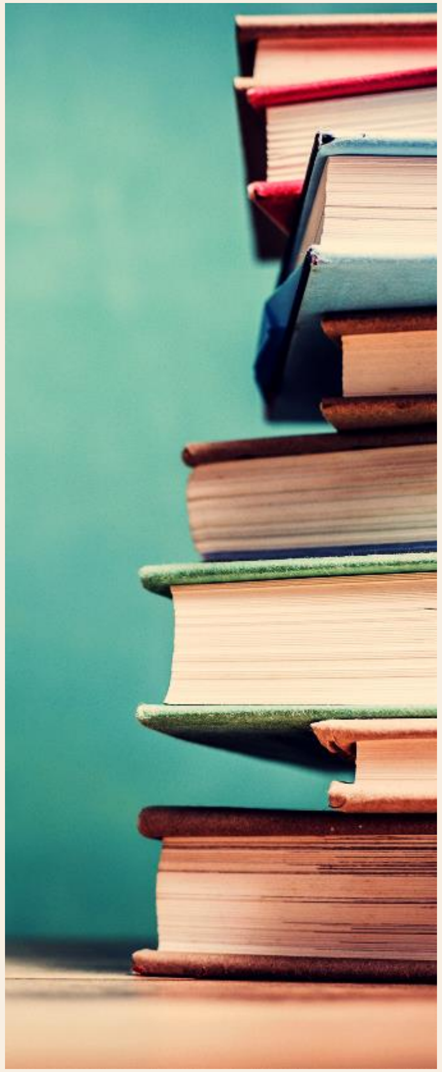
- PCA Overview
- Enhanced Clustering Results

Conclusion & Future Work

- Summary of Findings
- Next Steps and Recommendations

Q&A





Introduction

- Predictive maintenance helps anticipate equipment failures and optimize maintenance schedules.
- Matzka's Kaggle dataset, "Predictive Maintenance Dataset AI4I 2020," is a resource for developing predictive maintenance models.
- The dataset contains features related to machinery operations and failures.
- The goal is to build models to forecast failures and improve maintenance strategies across industries like manufacturing, transportation, and energy.

Project overview

- Analyze sensor data from industrial equipment for predictive maintenance.

Steps include:-

- Data preprocessing
- Exploratoray Data Analysis (EDA)
- Feature Engineering
- Model Building & Evaluation

✳ Use of machine learning algorithms like Random Forest, Gradient Boosting, and RNNs to detect failures.

Our Project at a glimpse

OUR DATASET

- AI4I 2020 dataset for predictive maintenance in manufacturing.
- Contains 10,000 rows and 14 columns.
- Key predictors: Product ID, Type, Air Temperature, Process Temperature, Rotational Speed, Torque, Tool Wear, Machine Failure.
- Includes binary targets for failure prediction.

TECHNIQUES & METHODOLOGIES

- Data Preprocessing: Cleaning, handling missing values, normalizing data, and outlier detection.
- EDA: Visualizing trends, identifying failure patterns.
- Feature Engineering: Extracting relevant features (statistical, time-based, dimensionality reduction).
- Predictive Modeling: Use of Random Forest, Gradient Boosting, SVM, and RNN for temporal data.
- Model Evaluation: Precision, recall, F1 score, AUC-ROC.

LIBRARIES USED

- Kaggle for dataset download.
- Pandas, Matplotlib, Seaborn for data processing and visualization.
- Scikit-learn for modeling (Random Forest, SVM), SMOTE for resampling.
- TensorFlow for deep learning models.

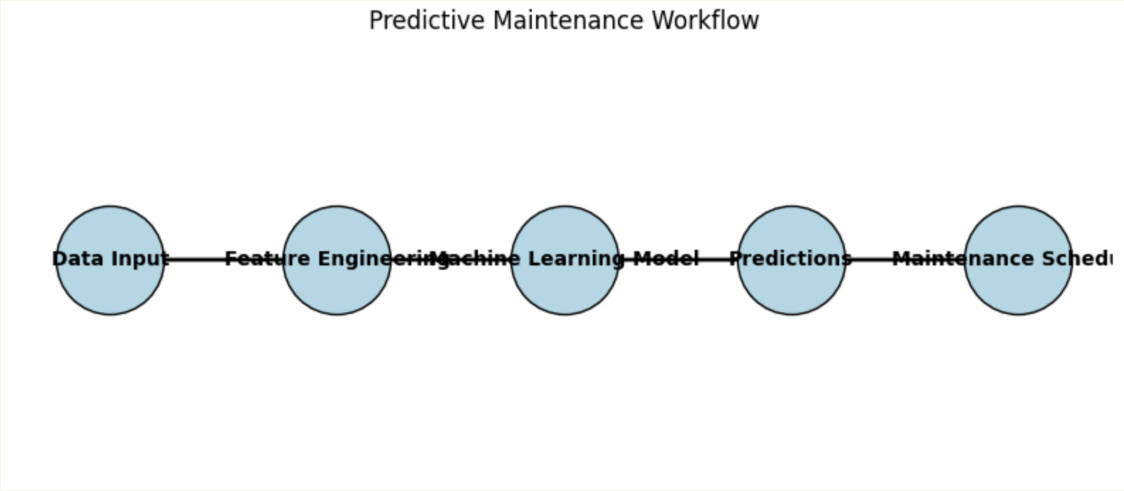
Introduction to Predictive Maintenance

Definition: Predictive maintenance uses data to predict equipment failures before they happen.

Benefits:

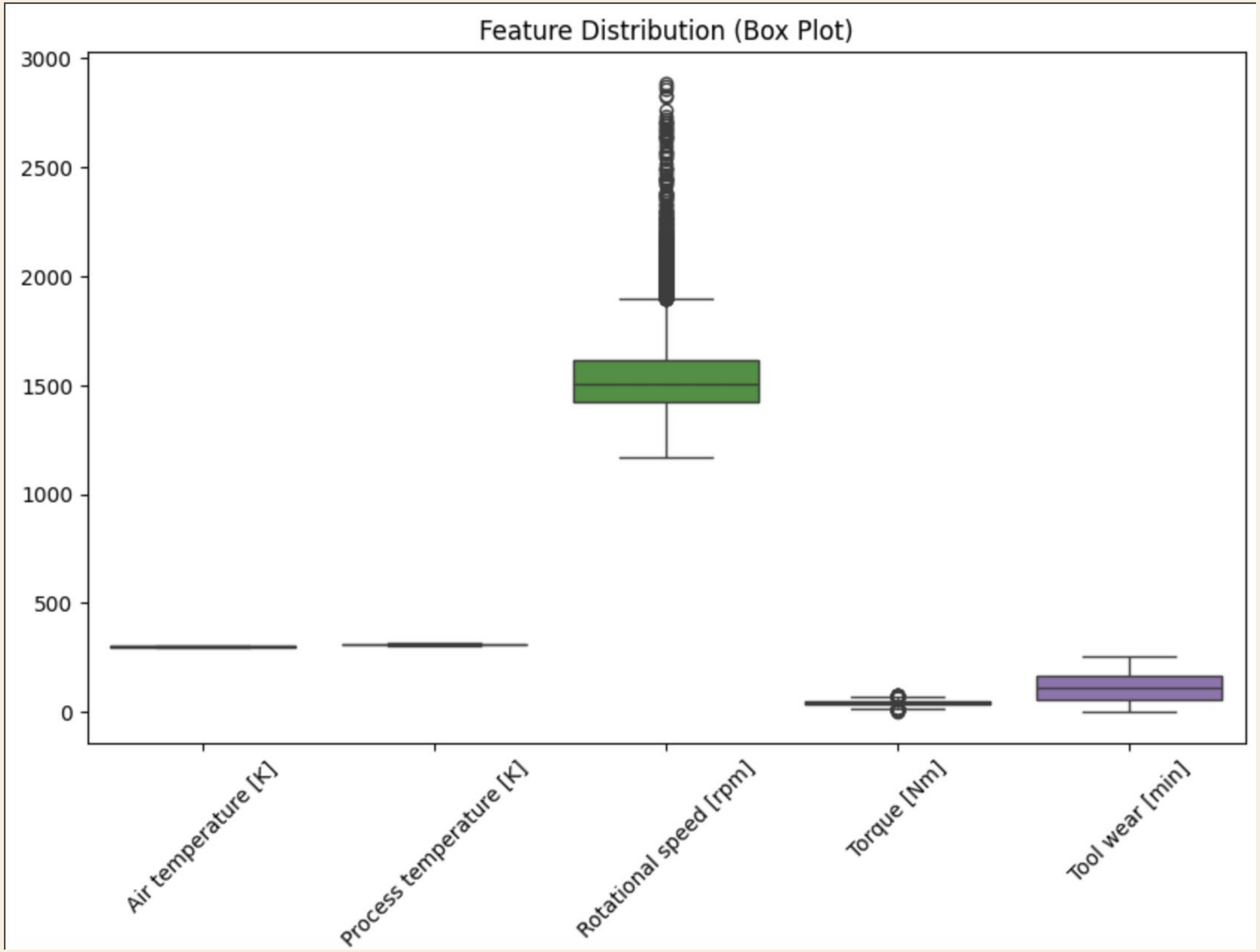
- Reduces downtime
- Enhances operational efficiency
- Decreases maintenance costs

Approach: Combining data analysis with machine learning to forecast potential failures.



Data Preparation

- **Data Splitting:** Divided into training and test sets for both classification and regression.
- **Feature Scaling:** Standardized features using StandardScaler for consistency.
- **Scaling Details:**
  - Mean = 0
  - Standard Deviation = 1



Feature Selection & Targets

- **Classifier Features:**
  - Air temperature [K]
  - Process temperature [K]
  - Rotational speed [rpm]
  - Torque [Nm]
- **Regression Features:** Same as classifier features
- **Targets:**
  - Classification: Machine failure (binary)
  - Regression: Tool wear [min]

Unique Values per Column:

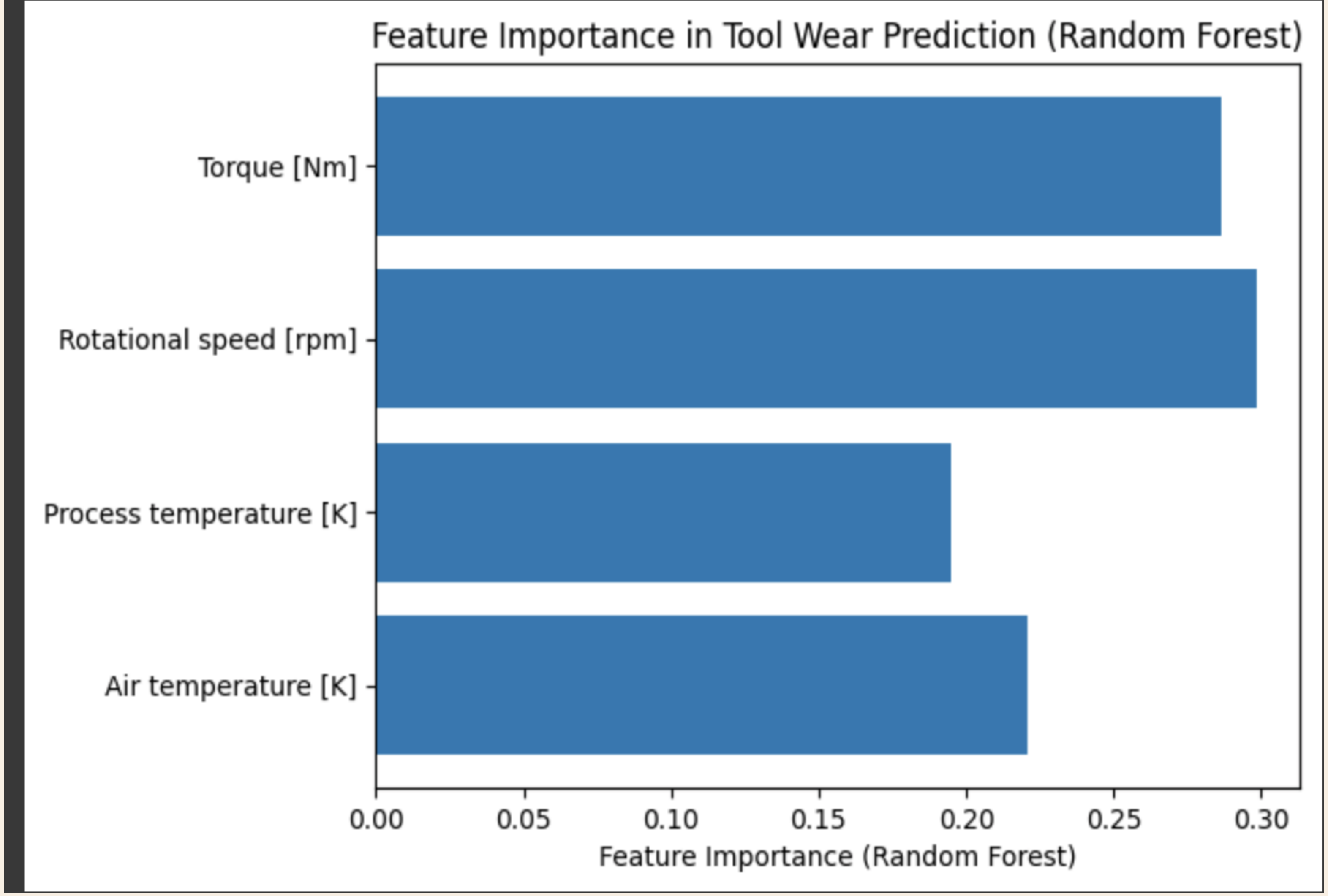
	0
UDI	10000
Product ID	10000
Type	3
Air temperature [K]	93
Process temperature [K]	82
Rotational speed [rpm]	941
Torque [Nm]	577
Tool wear [min]	246
Machine failure	2
TWF	2
HDF	2
PWF	2
OSF	2
RNF	2

dtype: int64

Model Training

- **Classification Model:** RandomForestClassifier
  - Robust to overfitting
  - Handles high-dimensional data well
- **Regression Model:** LinearRegression
  - Simple and interpretable
  - Provides predictions based on linear relationships
- **Random Forest Classifier:** Shows which features are most influential in predicting machine failure.
- Highest Importance: Torque [Nm]
- Lowest Importance: Air temperature [K]

```
Random Forest MSE: 0.999091366284536
Random Forest R^2: 0.05189462334906736
```



Predictions & Evaluation

- **Predictions Made:** Machine Failure: Classification results
- Tool Wear: Regression results
- **Regression Evaluation Metrics:** Mean Squared Error (MSE): Measures prediction accuracy
- R-squared (R^2): Indicates how well the model explains variability

```
Linear Regression MSE: 1.055039911963184
Linear Regression R^2: -0.001198736041068127
```

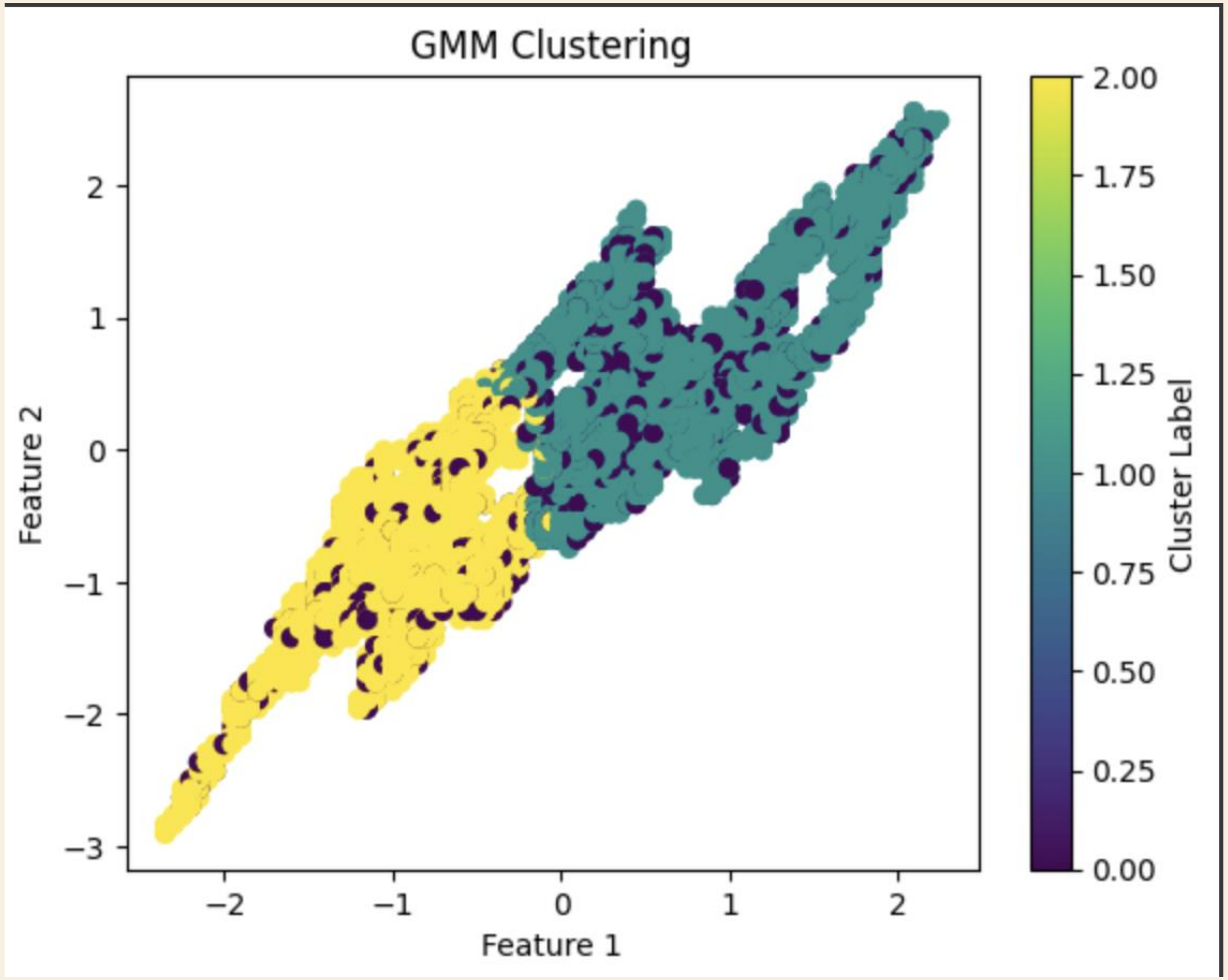
Clustering Analysis

Method Used: K-Means Clustering

- Number of Clusters: 3

- **Clustering Metrics:**
  - Silhouette Score: 0.2622 (Measures how well-separated clusters are)
  - Davies-Bouldin Score: 1.2437 (Evaluates clustering quality)

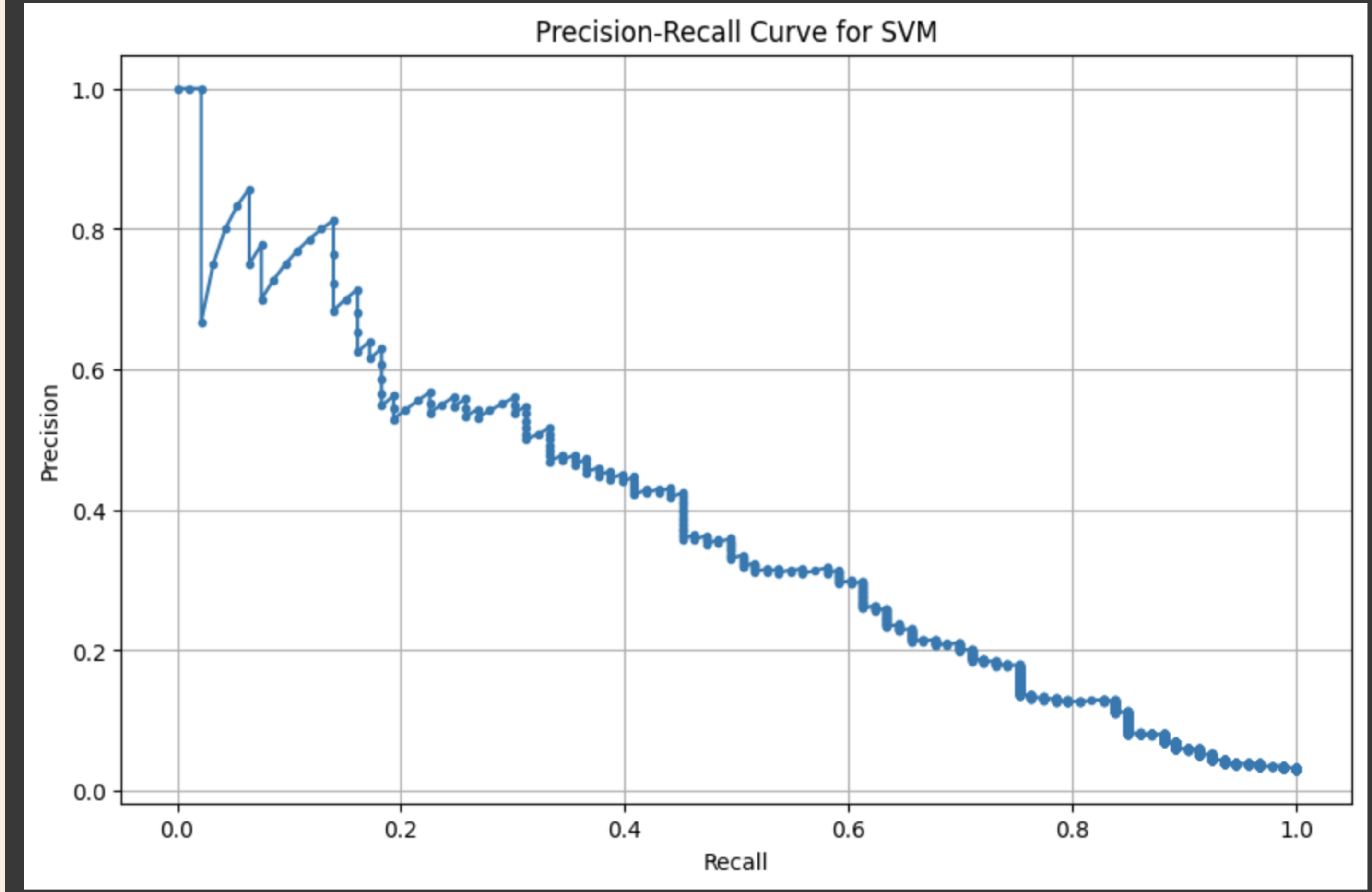
- We applied K-Means clustering to segment the data into three clusters. The Silhouette Score of 0.2622 indicates that clusters are not very distinct, while the Davies-Bouldin Score of 1.2437 suggests that the clusters are somewhat overlapping."



SVM Model

- VM ROC AUC score: 0.9765970904490828 Precision: .031. While 0.031 is very low, indicating that when the model predicts the positive class, it is rarely correct. This is a sign of poor performance for the positive class, despite the high ROC AUC score.

```
SVM ROC AUC score: 0.8733017447688375
```



	precision	recall	f1-score	support
0	0.97	1.00	0.98	2907
1	0.00	0.00	0.00	93
accuracy			0.97	3000
macro avg	0.48	0.50	0.49	3000
weighted avg	0.94	0.97	0.95	3000

PCA

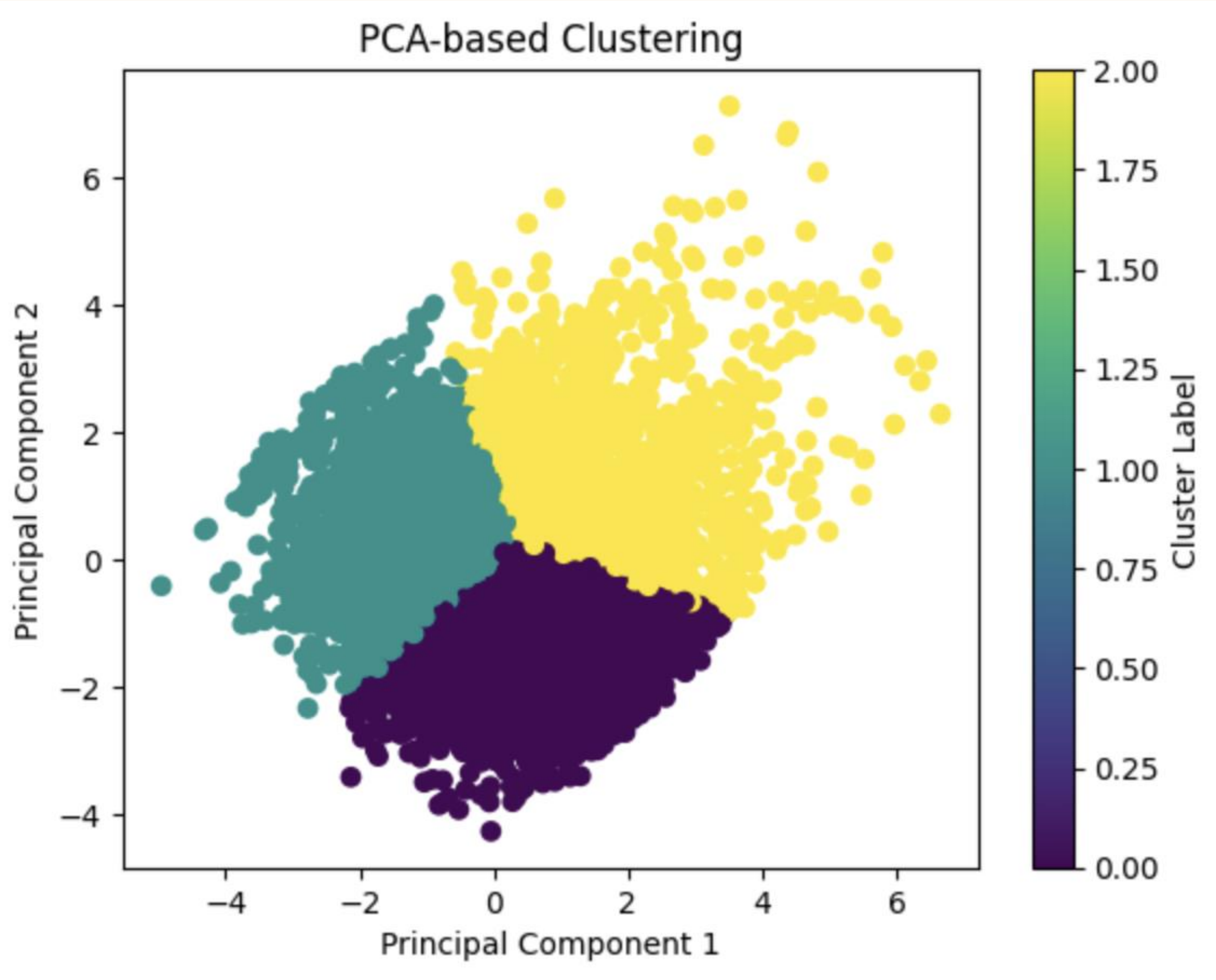
PCA FOR CLUSTERING IMPROVEMENT

- **PCA Application:** Reduced dimensions to 2 components.
- **Reconstruction Error Analysis:** Identifies anomalies based on error metrics.
- To improve clustering, we applied Principal Component Analysis (PCA) to reduce the data to two dimensions. This approach helps in visualizing clusters more clearly and allows us to analyze reconstruction errors to identify anomalies.

CLUSTERING METRICS POST-PCA

- **Silhouette Score after PCA:** [Insert Value]
- **Davies-Bouldin Score after PCA:** [Insert Value]
- **Impact of PCA:** Improved cluster separation or remained similar
- After applying PCA, we recalculated the clustering metrics. These updated scores help us understand whether dimensionality reduction improved cluster separation and the overall clustering quality

PCA



PCA-based Clustering
Principal Component 1
Principal Component 2
−4
−2
0
2
4
6
Cluster Label
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

Conclusion & Future Work

- In summary, our analysis highlights the effectiveness of predictive maintenance models and identifies some challenges in clustering. Moving forward, we plan to explore alternative clustering algorithms, enhance feature engineering, and experiment with additional dimensionality reduction techniques.

- **Future Directions:**
- Explore different clustering algorithms like DBSCAN or Gaussian Mixture Models.
- Enhance feature engineering for better clustering results.
- Experiment with further dimensionality reduction techniques.

Thanks for Tunning in!

Questions?